

Is a Data-Driven Approach still Better than Random Choice with Naive Bayes classifiers?

Piotr Szymański and Tomasz Kajdanowicz

Department of Computational Intelligence, Wrocław University of Technology,
Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland

Abstract. We study the performance of data-driven, a priori and random approaches to label space partitioning for multi-label classification with a Gaussian Naive Bayes classifier. Experiments were performed on 12 benchmark data sets and evaluated on 5 established measures of classification quality: micro and macro averaged F1 score, subset accuracy and Hamming loss. Data-driven methods are significantly better than an average run of the random baseline. In case of F1 scores and Subset Accuracy - data driven approaches were more likely to perform better than random approaches than otherwise in the worst case. There always exists a method that performs better than a priori methods in the worst case. The advantage of data-driven methods against a priori methods with a weak classifier is lesser than when tree classifiers are used.

Keywords: multi-label classification, label space clustering, data-driven classification

1 Introduction

In our recent work [11] we proposed a data-driven community detection approach to partition the label space for the multi-label classification as an alternative to random partitioning into equal subsets as performed by the random k -label sets method proposed by Tsoumakas et. al. [13]. The data-driven approach works as follows: we construct a label co-occurrence graph (both weighted and unweighted versions) based on training data and perform community detection to partition the label set. Then, each partition constitutes a label space for separate multi-label classification sub-problems. As a result, we obtain an ensemble of multi-label classifiers that jointly covers the whole label space. We consider a variety of approaches: modularity-maximizing techniques approximated by fast greedy and leading eigenvector methods, infomap, walktrap and label propagation algorithms. For comparison purposes we evaluate the binary relevance (BR) and label powerset (LP) - which we call a priori methods, as they a priori assume a total partitioning of the label space into singletons (BR) and lack of any partitioning (LP).

The variant of $RAkEL$ evaluated in this paper is an approach in which the label space is either partitioned into equal-sized subsets of labels. This approach is

called *RAkELd* - *RAkEL* distinct as the label sets are non-overlapping. *RAkELd* takes one parameter - the number of label sets to partition into k . We assumed that all partitions are equally probable and that the remainder of the label set smaller than k becomes the last element of the otherwise equally sized partition family.

In [11] we compared community detection methods to label space divisions against *RAkELd* and a priori methods on 12 benchmark datasets (*bibtex* [6], *delicious* [14], *tmc2007* [14], *enron* ([7]), *medical* [9], *scene* [1], *birds* [2], *Corel5k* [4], *Mediamill* [10], *emotions* [12], *yeast* [5], *genbase* [3]) over five evaluation measures with Classifier and Regression Trees (CART) as base classifiers. We discovered that data-driven approaches are more efficient and more likely to outperform *RAkELd* than binary relevance or label powerset is, in every evaluated measure. For all measures, apart from Hamming loss, data-driven approaches are significantly better than *RAkELd* ($\alpha = 0.05$), and at least one data-driven approach is more likely to outperform *RAkELd* than a priori methods in the case of *RAkELd*'s best performance. This has been the largest *RAkELd* evaluation published to date with 250 samplings per value for 10 values of *RAkELd* parameter k on 12 datasets published to date.

In this paper we extend our result and evaluate whether the same results hold if instead of using tree-based methods, we employ a weak and Gaussian Naive Bayesian classifier from the scikit-learn python package [8]. The experimental setup remains identical to the one presented in tree-based scheme, except for the change of base classifier. Bayesian classifiers remain of interest in many applications due to their low computational requirements.

We thus repeat the research questions we have asked in the case of tree-based classifiers, this time for Naive Bayes based classifiers:

- RH1:** Data-driven approach is significantly better than random ($\alpha = 0.05$)
- RH2:** Data-driven approach is more likely to outperform *RAkELd* than a priori methods
- RH3:** Data-driven approach is more likely to outperform *RAkELd* than a priori methods in the worst case
- RH4:** Data-driven approach is more likely to perform better than *RAkELd* in the worst case, than otherwise

2 Results

Micro-averaged F1 score. While a priori methods such as Binary Relevance and Label Powerset exhibit a higher median likelihood of outperforming *RAkELd* - we note that the highest mean likelihood is obtained by label propagation data-driven label space division on an unweighted label co-occurrence graph. Unweighted label propagation is also most likely to outperform *RAkELd* in the worst case. Thus we reject **RH2** and accept **RH3** and **RH4**. The best performing and recommended community detection method for micro-averaged F1 score - unweighted label propagation - is better than average performance of *RAkELd* with statistical significance, we thus accept **RH1**.

Macro-averaged F1 score. In case of macro averaged F1 score Label Powerset is the most likely to outperform RAKElD both in median and mean cases, while underperforms in the worst case. Label propagation data-driven label space division on an unweighted label co-occurrence graph is the most likely data-driven approach to outperform RAKElD - although other approaches also yield good results. Unweighted label propagation is also most likely to outperform RAKElD in the worst case. It is also better than an average run of RAKElD with statistical significance. Thus we accept **RH1**, reject **RH2** and accept **RH3** and **RH4**.

Subset Accuracy. In case of Subset Accuracy label propagation performed on an unweighted graph approach to dividing the labels space is the most resilient approach both in the worst case and in the average (mean/median) likelihood. The weighted version performs equally well in the worst case, so does unweighted infomap. As the worst case performance of three data-driven methods is greater than 0.5 we accept **RH4** for Subset Accuracy. While Label Powerset performs better than label propagation in case of the median/mean likelihood of being better than RAKElD - it performs worse by 12 pp. in the worst case. Thus while rejecting **RH2** and accepting **RH3** we still recommend using data-driven label propagation approach instead of Label Powerset. Label propagation performs better than RAKElD with statistical significance - we accept **RH1**.

Jaccard score. Among data-driven methods the label propagation performed on an unweighted graph approach to dividing the labels space is the most resilient approach both in the worst case and in the average (mean/median) likelihood. It is followed by infomap. While a priori methods are perform better in case of the median likelihood by 3 pp., they perform worse than data-driven methods in the mean and worst case. We thus confirm **RH2** and **RH3**. The worst case likelihood of data-driven methods outperforming RAKElD is not grater than 0.5 we thus reject **RH4**. Unweighted infomap performs better than the average run of RAKElD with statistical significance - we thus accept **RH1**.

Hamming Loss The data-driven methods that are most likely to outperform RAKElD are infomap and label propagation performed on a weighted label co-occurrence graph. We recommend using weighted infomap which is also most

| | FG | FGW | LE | LEW | WTW |
|-------------------|--------------|-------|--------------|--------------|--------------|
| Macro-averaged F1 | 0.068 | 0.37 | 0.054 | 0.37 | 0.37 |
| Micro-averaged F1 | 0.011 | 0.071 | 0.003 | 0.011 | 0.043 |
| Jaccard Score | 0.026 | 0.07 | 0.008 | 0.026 | 0.070 |

Table 1: P-values of data-driven methods performing better than an average run of RAKElD for each measure tested using non-parametric Friedman test with Rom’s post-hoc test. Only methods with p-values greater than $\alpha = 0.05$ are presented. All approaches not listed explicitly were significantly better than RAKElD in all measures.

resilient in the worst case, although much less resilient than the desired 0.5 likelihood of outperforming *RAkELd* in the worst case. As a result the case of Hamming Loss we confirm **RH2** and **RH3** but reject **RH4**. Weighted infomap perform significantly better than an average run of *RAkELd* - we accept **RH1**.

3 Conclusion and Outlook

We have examined the performance of data-driven, a priori and random approaches to label space partitioning for multi-label classification with a Gaussian Naive Bayes classifier. Experiments were performed on 12 benchmark data sets and evaluated on 5 established measures of classification quality. Table 12 summarizes our findings. Data-driven methods are significantly better than an average *RAkELd* run that had not undergone parameter estimation - i.e. when results are compared against the mean result of all evaluated *RAkELd* parameter values. When compared against the likelihood of outperforming a *RAkELd* in the evaluated parameter space - in case of F1 scores and Subset Accuracy - data driven approaches were more likely to perform better than *RAkELd* than otherwise in the worst case. There always exists a method that performs better than a priori methods in the worst case.

Data driven methods perform better than a priori methods in the mean likelihood but worse in median when it comes to micro-averaged F1 and Subset Accuracy. This can be attributed to differences in how likelihoods per data set distribute - data-driven methods perform better in worst case, but are also less likely to be always better than *RAkELd* as opposed to a priori methods. The advantage of data-driven methods against a priori methods with a weak classifier is lesser than when tree classifiers are used. The authors acknowledge support from the National Science Centre research projects decision no. 2016/21/N/ST6/02382 and 2016/21/D/ST6/02948.

| | Micro-averaged F1 | Macro-averaged F1 | Subset accuracy | Ac-Jaccard Similarity | Hamming Loss |
|----------------------|-------------------------------|------------------------------|------------------------------|------------------------------|------------------|
| RH1 | Yes | Yes | Yes | Yes | Yes |
| RH2 | Undecided | No | No | Undecided | Yes |
| RH3 | Yes | Yes | Yes | Yes | Yes |
| RH4 | Yes | Yes | Yes | No | No |
| Recommended approach | data-driven label propagation | Unweighted label propagation | Unweighted label propagation | Unweighted label propagation | Weighted infomap |

Table 12: The summary of evaluated hypotheses and proposed recommendations of this paper

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. *Pattern Recognition* 37(9), 1757–1771 (Sep 2004), <http://www.sciencedirect.com/science/article/pii/S0031320304001074>
2. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J.K., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America* 131(6), 4640–4650 (2012), <http://scitation.aip.org/content/asa/journal/jasa/131/6/10.1121/1.4707424>
3. Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I.: Protein Classification with Multiple Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 448–456 (2005), <http://www.springerlink.com/index/P662542G78792762.pdf>
4. Duygulu, P., Barnard, K., Freitas, J.F.G.d., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Proceedings of the 7th European Conference on Computer Vision-Part IV*. p. 97–112. *ECCV '02*, Springer-Verlag, London, UK, UK (2002), <http://dl.acm.org/citation.cfm?id=645318.649254>
5. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: *In Advances in Neural Information Processing Systems 14*. pp. 681–687. MIT Press (2001)
6. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: *In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge* (2008)
7. Kliment, B., Yang, Y.: The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004* pp. 217–226 (2004), <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.1645&rep=rep1&type=pdf>
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011)
9. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. *Machine Learning* 85(3), 333–359 (Dec 2011), <http://link.springer.com/article/10.1007/s10994-011-5256-5>
10. Snoek, C.G.M., Worring, M., Gemert, J.C.V., Geusebroek, J.m., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: *In Proceedings of the ACM International Conference on Multimedia*. p. 421–430. ACM Press (2006)
11. Szymanski, P., Kajdanowicz, T., Kersting, K.: How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy* 18(8), 282 (2016), <http://dx.doi.org/10.3390/e18080282>
12. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: *ISMIR*. vol. 8, pp. 325–330 (2008)
13. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4701, pp. 406–417. Springer (2007)
14. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*. p. 30–44 (2008)

| | Minimum | Median | Mean | Std |
|------------------------------|-----------------|-----------------|-----------------|----------|
| BR | 0.369565 | 1.000000 | 0.796143 | 0.283117 |
| LP | 0.369565 | 0.999076 | 0.789076 | 0.294146 |
| fastgreedy | 0.263556 | 0.781778 | 0.737634 | 0.232243 |
| fastgreedy-weighted | 0.322667 | 0.601848 | 0.633698 | 0.160196 |
| infomap | 0.448000 | 0.869778 | 0.817113 | 0.194957 |
| infomap-weighted | 0.091556 | 0.797333 | 0.705199 | 0.299230 |
| label_propagation | 0.529778 | 0.908000 | 0.843744 | 0.172125 |
| label_propagation-weighted | 0.317778 | 0.662356 | 0.703653 | 0.243097 |
| leading_eigenvector | 0.302667 | 0.829778 | 0.748593 | 0.250929 |
| leading_eigenvector-weighted | 0.341778 | 0.632063 | 0.684237 | 0.185325 |
| walktrap | 0.321333 | 0.717391 | 0.719968 | 0.246686 |
| walktrap-weighted | 0.239556 | 0.600889 | 0.632683 | 0.221396 |

Table 2: Likelihood of performing better than *RAkELd* in Micro-averaged F1 score of every method for each data set

| | BR | LP | FG | FGW | IN | INW | LPG | LPGW | LE | LEW | WT | WTW |
|-------------|------|-------|------|------|------|------|------|------|------|------|------|-------|
| Corel5k | 0.39 | 0.37 | 0.85 | 0.79 | 0.87 | 0.09 | 0.99 | 0.32 | 0.9 | 0.91 | 0.43 | 0.68 |
| bibtex | 0 | 0 | 0.26 | 0.32 | 0.45 | 0.3 | 0.53 | 0.34 | 0.30 | 0.34 | 0.32 | 0.24 |
| birds | 0 | 0.999 | 0.62 | 0.6 | 0.66 | 0.66 | 0.66 | 0.66 | 0.79 | 0.62 | 0.95 | 0.34 |
| delicious | 0 | 0 | 0.78 | 0.59 | 0.87 | 0.59 | 0.87 | 0.62 | 0.83 | 0.72 | 0.54 | 0.58 |
| emotions | 0.43 | 0.37 | 0 | 0.52 | 0 | 0 | 0 | 0.57 | 0 | 0.52 | 0 | 0.57 |
| enron | 0.98 | 0.98 | 0.94 | 0.88 | 0.93 | 0.93 | 0.93 | 0.93 | 0 | 0.99 | 0.79 | 0.997 |
| mediamill | 0 | 0 | 0.55 | 0.65 | 0.91 | 0.8 | 0.91 | 0.91 | 0.45 | 0.69 | 0.68 | 0.6 |
| medical | 0 | 0 | 0.51 | 0.58 | 0.51 | 0.60 | 0.60 | 0.60 | 0.41 | 0.59 | 0.51 | 0.60 |
| scene | 0.37 | 0.37 | 0.72 | 0.63 | 0.80 | 0.80 | 0.80 | 0.80 | 0.72 | 0.63 | 0.72 | 0.63 |
| tmc2007-500 | 0 | 0 | 0.89 | 0.55 | 0 | 0 | 0 | 0 | 0.85 | 0.63 | 0 | 0.89 |
| yeast | 0.58 | 0.59 | 0.99 | 0.85 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.88 | 0.99 | 0.83 |

Table 3: Likelihood of performing better than *RAkELd* in Micro-averaged F1 score of every method for each data set. BR - Binary Relevance, LP - Label Powerset, FG - fastgreedy, FGW - fastgreedy weighted, IN - infomap, INW - infomap weighted, LPG - label propagation, LPGW - label propagation weighted, LE - leading eigenvector, LEW - leading eigenvector weighted, WT - walktrap, WTW - walktrap weighted.

| | Minimum | Median | Mean | Std |
|------------------------------|-----------------|-----------------|-----------------|----------|
| BR | 0.456522 | 1.000000 | 0.868708 | 0.222246 |
| LP | 0.434783 | 1.000000 | 0.850310 | 0.227355 |
| fastgreedy | 0.376444 | 0.836000 | 0.799503 | 0.210402 |
| fastgreedy-weighted | 0.378222 | 0.753333 | 0.679727 | 0.175535 |
| infomap | 0.519630 | 0.806861 | 0.810572 | 0.164820 |
| infomap-weighted | 0.188444 | 0.739130 | 0.728628 | 0.247947 |
| label_propagation | 0.519630 | 0.878667 | 0.841961 | 0.163304 |
| label_propagation-weighted | 0.500000 | 0.739130 | 0.751203 | 0.186984 |
| leading_eigenvector | 0.367111 | 0.806861 | 0.746465 | 0.232450 |
| leading_eigenvector-weighted | 0.358667 | 0.832457 | 0.722748 | 0.215705 |
| walktrap | 0.253778 | 0.877333 | 0.789586 | 0.225409 |
| walktrap-weighted | 0.302222 | 0.800444 | 0.745813 | 0.235022 |

Table 4: Likelihood of performing better than *RAkELd* in Macro-averaged F1 score of every method for each data set

| | BR | LP | FG | FGW | IN | INW | LPG | LPGW | LE | LEW | WT | WTW |
|-------------|------|-------|-------|------|------|------|-------|------|------|------|------|------|
| Corel5k | 0.94 | 0.78 | 0.37 | 0.37 | 0.89 | 0.18 | 0.997 | 0.76 | 0.36 | 0.36 | 0.25 | 0.3 |
| bibtex | 1.0 | 1.0 | 0.53 | 0.57 | 0.67 | 0.52 | 0.88 | 0.55 | 0.52 | 0.61 | 0.6 | 0.47 |
| birds | 1.0 | 1.0 | 0.98 | 0.84 | 0.52 | 0.52 | 0.52 | 0.52 | 0.99 | 0.96 | 0.97 | 0.97 |
| delicious | 1.0 | 1.0 | 1.0 | 0.79 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.85 | 0.88 | 0.97 |
| emotions | 0.46 | 0.46 | 0.93 | 0.52 | 0.93 | 0.93 | 0.93 | 0.5 | 0.93 | 0.52 | 0.93 | 0.5 |
| enron | 1.0 | 0.998 | 0.986 | 0.89 | 0.66 | 0.66 | 0.66 | 0.66 | 0.88 | 0.89 | 0.99 | 0.91 |
| mediamill | 1.0 | 1.0 | 0.84 | 0.75 | 0.99 | 0.91 | 0.99 | 0.99 | 0.76 | 0.84 | 0.89 | 0.8 |
| medical | 1.0 | 1.0 | 0.7 | 0.45 | 0.7 | 0.74 | 0.74 | 0.74 | 0.39 | 0.45 | 0.70 | 0.74 |
| scene | 0.46 | 0.43 | 0.65 | 0.65 | 0.74 | 0.74 | 0.74 | 0.74 | 0.65 | 0.65 | 0.65 | 0.65 |
| tmc2007-500 | 1.0 | 1.0 | 0.99 | 0.8 | 1.0 | 1.0 | 1.0 | 1.0 | 0.92 | 0.83 | 1.0 | 0.98 |
| yeast | 0.7 | 0.68 | 0.8 | 0.83 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 1.0 | 0.81 | 0.91 |

Table 5: Likelihood of performing better than *RAkELd* in Macro-averaged F1 score of every method for each data set. BR - Binary Relevance, LP - Label Powerset, FG - fastgreedy, FGW - fastgreedy weighted, IN - infomap, INW - infomap weighted, LPG - label propagation, LPGW - label propagation weighted, LE - leading eigenvector, LEW - leading eigenvector weighted, WT - walktrap, WTW - walktrap weighted.

| | Minimum | Median | Mean | Std |
|------------------------------|-----------------|-----------------|-----------------|----------|
| BR | 0.217391 | 0.886667 | 0.777640 | 0.285316 |
| LP | 0.413043 | 1.000000 | 0.924946 | 0.174772 |
| fastgreedy | 0.028637 | 0.585333 | 0.621030 | 0.304067 |
| fastgreedy-weighted | 0.007852 | 0.586728 | 0.512003 | 0.225171 |
| infomap | 0.429000 | 0.978500 | 0.887924 | 0.203588 |
| infomap-weighted | 0.533487 | 0.934783 | 0.831424 | 0.195409 |
| label_propagation | 0.533487 | 0.998222 | 0.912394 | 0.165066 |
| label_propagation-weighted | 0.533487 | 0.934783 | 0.834437 | 0.180916 |
| leading_eigenvector | 0.000000 | 0.644000 | 0.604389 | 0.355451 |
| leading_eigenvector-weighted | 0.000000 | 0.568988 | 0.499787 | 0.304284 |
| walktrap | 0.133487 | 0.600000 | 0.625201 | 0.295569 |
| walktrap-weighted | 0.000000 | 0.608696 | 0.499824 | 0.331589 |

Table 6: Likelihood of performing better than RAKElD in Subset Accuracy of every method for each data set

| | BR | LP | FG | FGW | IN | INW | LPG | LPGW | LE | LEW | WT | WTW |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|------|
| Corel5k | 0.34 | 0.87 | 0.59 | 0.68 | 0.99 | 0.59 | 0.998 | 0.83 | 0.0 | 0.0 | 0.34 | 0.0 |
| bibtex | 0.89 | 1.0 | 0.37 | 0.69 | 0.96 | 0.61 | 0.998 | 0.7 | 0.64 | 0.73 | 0.37 | 0.31 |
| birds | 0.996 | 0.997 | 0.029 | 0.007 | 0.53 | 0.53 | 0.53 | 0.53 | 0.09 | 0.0 | 0.13 | 0.03 |
| delicious | 1.0 | 1.0 | 0.997 | 0.63 | 1.0 | 0.999 | 1.0 | 1.0 | 1.0 | 0.79 | 0.79 | 0.92 |
| emotions | 0.21 | 0.41 | 1.0 | 0.48 | 1.0 | 1.0 | 1.0 | 0.61 | 1.0 | 0.48 | 1.0 | 0.61 |
| enron | 0.86 | 1.0 | 0.58 | 0.57 | 0.98 | 0.98 | 0.98 | 0.98 | 0.79 | 0.67 | 0.6 | 0.65 |
| mediamill | 1.0 | 1.0 | 0.45 | 0.29 | 0.96 | 0.87 | 0.96 | 0.96 | 0.41 | 0.38 | 0.57 | 0.28 |
| medical | 1.0 | 1.0 | 0.43 | 0.64 | 0.43 | 0.64 | 0.64 | 0.64 | 0.33 | 0.65 | 0.43 | 0.64 |
| scene | 0.63 | 0.93 | 0.63 | 0.3 | 0.93 | 0.93 | 0.93 | 0.93 | 0.63 | 0.3 | 0.63 | 0.3 |
| tmc2007-500 | 1.0 | 1.0 | 0.75 | 0.59 | 1.0 | 1.0 | 1.0 | 1.0 | 0.75 | 0.57 | 1.0 | 0.87 |
| yeast | 0.62 | 0.96 | 0.999 | 0.76 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.92 | 0.999 | 0.88 |

Table 7: Likelihood of performing better than RAKElD in Subset Accuracy of every method for each data set. BR - Binary Relevance, LP - Label Powerset, FG - fastgreedy, FGW - fastgreedy weighted, IN - infomap, INW - infomap weighted, LPG - label propagation, LPGW - label propagation weighted, LE - leading eigenvector, LEW - leading eigenvector weighted, WT - walktrap, WTW - walktrap weighted.

| | Minimum | Median | Mean | Std |
|------------------------------|-----------------|-----------------|----------------|----------|
| BR | 0.326087 | 1.000000 | 0.784597 | 0.303331 |
| LP | 0.369565 | 1.000000 | 0.847350 | 0.240611 |
| fastgreedy | 0.183372 | 0.756000 | 0.674557 | 0.274675 |
| fastgreedy-weighted | 0.177367 | 0.586957 | 0.591697 | 0.194144 |
| infomap | 0.411085 | 0.925333 | 0.831944 | 0.218665 |
| infomap-weighted | 0.053778 | 0.804889 | 0.686328 | 0.327207 |
| label_propagation | 0.411085 | 0.974500 | 0.86552 | 0.203504 |
| label_propagation-weighted | 0.239111 | 0.630435 | 0.689132 | 0.281967 |
| leading_eigenvector | 0.308000 | 0.777333 | 0.693396 | 0.272005 |
| leading_eigenvector-weighted | 0.116859 | 0.653745 | 0.624674 | 0.222935 |
| walktrap | 0.359556 | 0.696444 | 0.668188 | 0.252658 |
| walktrap-weighted | 0.080370 | 0.586957 | 0.580375 | 0.244502 |

Table 8: Likelihood of performing better than *RAkELd* in Jaccard Similarity of every method for each data set

| | BR | LP | FG | FGW | IN | INW | LPG | LPGW | LE | LEW | WT | WTW |
|-------------|-------|-------|-------|------|-------|-------|-------|-------|-------|------|-------|------|
| Corel5k | 0.35 | 0.47 | 0.76 | 0.8 | 0.9 | 0.05 | 0.996 | 0.24 | 0.78 | 0.83 | 0.43 | 0.57 |
| bibtex | 1.0 | 1.0 | 0.31 | 0.42 | 0.86 | 0.40 | 0.99 | 0.45 | 0.42 | 0.45 | 0.36 | 0.28 |
| birds | 1.0 | 0.999 | 0.18 | 0.18 | 0.41 | 0.41 | 0.41 | 0.41 | 0.32 | 0.12 | 0.45 | 0.08 |
| delicious | 1.0 | 1.0 | 0.7 | 0.53 | 0.77 | 0.44 | 0.77 | 0.47 | 0.74 | 0.7 | 0.49 | 0.49 |
| emotions | 0.33 | 0.37 | 0.98 | 0.52 | 0.98 | 0.98 | 0.98 | 0.63 | 0.98 | 0.52 | 0.98 | 0.63 |
| enron | 0.993 | 1.0 | 0.84 | 0.82 | 0.97 | 0.97 | 0.97 | 0.97 | 0.999 | 0.87 | 0.74 | 0.88 |
| mediamill | 1.0 | 1.0 | 0.54 | 0.65 | 0.93 | 0.80 | 0.93 | 0.93 | 0.44 | 0.68 | 0.7 | 0.6 |
| medical | 1.0 | 1.0 | 0.41 | 0.55 | 0.41 | 0.55 | 0.55 | 0.55 | 0.31 | 0.56 | 0.41 | 0.55 |
| scene | 0.39 | 0.85 | 0.80 | 0.59 | 0.93 | 0.93 | 0.93 | 0.93 | 0.8 | 0.59 | 0.8 | 0.59 |
| tmc2007-500 | 1.0 | 1.0 | 0.89 | 0.6 | 1.0 | 1.0 | 1.0 | 1.0 | 0.85 | 0.65 | 1.0 | 0.9 |
| yeast | 0.57 | 0.63 | 0.994 | 0.85 | 0.994 | 0.994 | 0.994 | 0.994 | 0.994 | 0.91 | 0.994 | 0.82 |

Table 9: Likelihood of performing better than *RAkELd* in Jaccard Similarity of every method for each data set

| | Minimum | Median | Mean | Std |
|------------------------------|-----------------|-----------------|-----------------|----------|
| BR | 0.110667 | 0.558538 | 0.579872 | 0.376954 |
| LP | 0.080889 | 0.652174 | 0.592830 | 0.379345 |
| fastgreedy | 0.208889 | 0.418222 | 0.513625 | 0.276367 |
| fastgreedy-weighted | 0.111111 | 0.260870 | 0.302981 | 0.223065 |
| infomap | 0.112889 | 0.735111 | 0.684758 | 0.292563 |
| infomap-weighted | 0.204889 | 0.847826 | 0.727799 | 0.291282 |
| label_propagation | 0.111111 | 0.735111 | 0.684971 | 0.312656 |
| label_propagation-weighted | 0.237778 | 0.735111 | 0.714660 | 0.237049 |
| leading_eigenvector | 0.121333 | 0.498029 | 0.552381 | 0.315482 |
| leading_eigenvector-weighted | 0.111111 | 0.260870 | 0.337735 | 0.227415 |
| walktrap | 0.111111 | 0.418667 | 0.541611 | 0.331449 |
| walktrap-weighted | 0.094226 | 0.328113 | 0.387505 | 0.228658 |

Table 10: Likelihood of performing better than *RAkELd* in Hamming Loss of every method for each data set

| | BR | LP | FG | FGW | IN | INW | LPG | LPGW | LE | LEW | WT | WTW |
|-------------|-------|-------|------|------|------|------|------|------|------|------|------|------|
| Corel5k | 0.11 | 0.15 | 0.42 | 0.35 | 0.33 | 0.96 | 0.23 | 0.86 | 0.21 | 0.22 | 0.42 | 0.42 |
| bibtex | 0.11 | 0.08 | 0.31 | 0.24 | 0.11 | 0.20 | 0.11 | 0.24 | 0.27 | 0.24 | 0.17 | 0.26 |
| birds | 1.0 | 0.99 | 0.27 | 0.16 | 0.7 | 0.7 | 0.7 | 0.7 | 0.37 | 0.2 | 0.41 | 0.09 |
| delicious | 0.11 | 0.11 | 0.34 | 0.11 | 0.36 | 0.24 | 0.36 | 0.39 | 0.41 | 0.11 | 0.11 | 0.2 |
| emotions | 0.43 | 0.30 | 1.0 | 0.28 | 1.0 | 1.0 | 1.0 | 0.54 | 1.0 | 0.28 | 1.0 | 0.54 |
| enron | 0.40 | 0.69 | 0.31 | 0.3 | 0.73 | 0.73 | 0.73 | 0.73 | 0.94 | 0.57 | 0.27 | 0.43 |
| mediamill | 0.998 | 0.999 | 0.21 | 0.23 | 0.74 | 0.51 | 0.74 | 0.74 | 0.12 | 0.31 | 0.35 | 0.22 |
| medical | 1.0 | 1.0 | 0.77 | 0.94 | 0.77 | 0.88 | 0.88 | 0.88 | 0.79 | 0.93 | 0.77 | 0.88 |
| scene | 0.65 | 0.65 | 0.52 | 0.26 | 0.85 | 0.85 | 0.85 | 0.85 | 0.52 | 0.26 | 0.52 | 0.26 |
| tmc2007-500 | 1.0 | 1.0 | 0.57 | 0.17 | 1.0 | 1.0 | 1.0 | 1.0 | 0.5 | 0.26 | 1.0 | 0.64 |
| yeast | 0.56 | 0.55 | 0.94 | 0.3 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.33 | 0.94 | 0.33 |

Table 11: Likelihood of performing better than *RAkELd* in Hamming Loss of every method for each data set. BR - Binary Relevance, LP - Label Powerset, FG - fastgreedy, FGW - fastgreedy weighted, IN - infomap, INW - infomap weighted, LPG - label propagation, LPGW - label propagation weighted, LE - leading eigenvector, LEW - leading eigenvector weighted, WT - walktrap, WTW - walktrap weighted.